ABSTRACT OF THE DISCLOSURE

The present invention is a method and apparatus to reduce latency in accessing a memory from a bus. The apparatus comprises a pre-fetcher and a cache controller. The pre-fetcher pre-fetches a plurality of data from the memory to a cache queue in response to a request. The cache controller is coupled to the cache queue and the pre-fetcher to deliver the pre-fetched data from the cache queue to the bus in a pipeline chain independently of the memory.